Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed by Somatic Mutations in 7,000 Cancer Genomes

Qiancheng Shen,[1,8] Feixiong Cheng,[2,3,6,7,8] Huili Song,[1] Weiqiang Lu,[4] Junfei Zhao,[3] Xiaoli An,[1] Mingyao Liu,[4] Guoqiang Chen,[1] Zhongming Zhao,[5,*] and Jian Zhang[1,*]

The allosteric regulation triggering the protein's functional activity via conformational changes is an intrinsic function of protein under many physiological and pathological conditions, including cancer. Identification of the biological effects of specific somatic variants on allosteric proteins and the phenotypes that they alter during tumor initiation and progression is a central challenge for cancer genomes in the post-genomic era. Here, we mapped more than 47,000 somatic missense mutations observed in approximately 7,000 tumor-normal matched samples across 33 cancer types into protein allosteric sites to prioritize the mutated allosteric proteins and we tested our prediction in cancer cell lines. We found that the deleterious mutations identified in cancer genomes were more significantly enriched at protein allosteric sites than tolerated mutations, suggesting a critical role for protein allosteric variants in cancer. Next, we developed a statistical approach, namely AlloDriver, and further identified 15 potential mutated allosteric proteins during pan-cancer and individual cancer-type analyses. More importantly, we experimentally confirmed that p.Pro360Ala on PDE10A played a potential oncogenic role in mediating tumorigenesis in non-small cell lung cancer (NSCLC). In summary, these findings shed light on the role of allosteric regulation during tumorigenesis and provide a useful tool for the timely development of targeted cancer therapies.

Introduction

Cancer is a major public health problem and is currently the second leading cause of death in the United States.[1] Recently, next-generation sequencing (NGS) technology, including whole-exome and whole-genome sequencing, has helped investigators uncover massive amounts of somatic alterations in cancer genomes in several large-scale projects, such as The Cancer Genome Atlas (TCGA)[2] and International Cancer Genome Consortium (ICGC).[3] Furthermore, these studies demonstrated that most cancers harbor only a few significantly mutated genes (SMGs) in each cancer genome and that many cancer-associated genes are mutated in a small number of individuals.[4] For instance, a recent study has suggested that a typical tumor genome contains two to eight driver gene mutations.[4] Accordingly, the majority of the remaining somatic alterations are called "passenger mutations," which have no biologically relevant effects on tumor fitness and progression.[5] The systematic elucidation of the functional consequences of somatic mutations in cancer is a big challenge in the era of the human post-genome projects.[6] Identifying the variants altering protein function is a promising strategy for deciphering the biological consequences of somatic mutations during tumorigenesis and would provide novel targets for the development of targeted cancer therapies.[7]

Receptors are a class of proteins with dual roles in the recognition of a drug or environmental factors and the transduction of these stimuli into cellular responses. Although most studies on receptor function have focused on how ligands modulate receptor signaling pathways by binding to orthosteric sites, receptor conformation and signal transduction can also be regulated by ligands acting on unique allosteric sites.[8] Topographically, an allosteric site is an area of a protein distinct from the orthosteric site that can regulate the protein's functional activity via conformational changes induced by the binding of allosteric ligands.[9] Pathological orthosteric (at the substrate-binding site) and allosteric (at the allosteric site) events can deregulate a protein, trapping it in either its active or inactive conformation.[10] Furthermore, uncontrolled protein activity typically leads to disease.[10] Additionally, cells have various molecular structures that form complex, dynamic, and plastic networks.[11] Under the molecular network framework, somatic mutations may alter network architecture by affecting nodes (i.e., proteins), edges (i.e., protein interactions), or both within a network or by changing the biochemical properties of nodes.[12–14] The large amount of NGS data generated from cancer

[1]Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of National Ministry of Education, Shanghai Key Laboratory of Tumor Microenvironment and Inflammation, Shanghai Jiao-Tong University School of Medicine, Shanghai 200025, China; [2]State Key Laboratory of Biotherapy/Collaborative Innovation Center for Biotherapy, West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, Sichuan, China; [3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA; [4]Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China; [5]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[6]Present address: Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[7]Present address: Center for Complex Networks Research, Northeastern University, Boston, MA 02115, USA
[8]These authors contributed equally to this work
*Correspondence: zhongming.zhao@uth.tmc.edu (Z.Z.), jian.zhang@sjtu.edu.cn (J.Z.)
http://dx.doi.org/10.1016/j.ajhg.2016.09.020.
© 2017 American Society of Human Genetics.

genome projects, such as TCGA and ICGC, provide us with an unprecedented opportunity to systematically examine allosteric regulation related to tumor initiation and progression. So far, to the best knowledge of the authors, there has been no systematic investigation of the large-scale allosteric regulation perturbed by somatic mutations in cancer.

In this study, we employed an integrative genomics workflow to systematically investigate cancer allosteric regulations perturbed by somatic variants at allosteric sites. We manually constructed a catalog of allosteric proteins curated from the literature based on our previous studies.[15,16] We found that the deleterious mutations identified in cancer genomes were more significantly enriched at protein allosteric sites than tolerated mutations, suggesting a critical role for protein allosteric variants in tumor initiation and progression. Next, we developed a statistical approach, namely AlloDriver, to prioritize potentially functional mutations in cancer via altering protein allosteric regulation in both pan-cancer and individual cancer types. In a case study, we tested the results predicted by the model experimentally. Specifically, we mapped more than 47,000 somatic missense mutations generated from approximately 7,000 tumor-normal matched samples to protein allosteric sites derived from protein three-dimensional (3D) structures and our large-scale, manually curated experimental data. We identified 15 potential significantly mutated proteins harboring enriched somatic variants via altering protein allosteric regulation during pan-cancer and individual cancer type analyses using AlloDriver. Then, we experimentally verified the functional role of p.Pro360Ala on PDE10A using non-small cell lung cancer (NSCLC) as a case study. In summary, this study provides insights into cancer allosteric regulation perturbations altered by somatic variants and provides a powerful tool for the development of novel targeted cancer therapies.

Material and Methods

Construction of a Catalog of Allosteric Proteins

The comprehensive allosteric protein catalog was obtained from the AlloSteric Database (ASD) constructed by our group,[16] which provides a versatile resource of the well-established allosteric macromolecules and ligands found since 1901. The version of the ASD includes 1,286 allosteric proteins distributed across 181 different species covering prokaryotes and eukaryotes and 22,008 allosteric ligands. In our ASD curation, proteins with at least three cases of experimental evidences in crystal structure complex or biochemistry (such as site-directed mutagenesis, cooperativity of kinetic effect from two ligands, and uncompetitive binding assay with chromatography, etc.) supporting their functional change elicited by ligand binding at a site that was topographically distinct from the orthosteric functional site were considered as allosteric proteins and deposited into the ASD. Among these allosteric proteins, 574 proteins belong to human species (Table S1), including 74 experimentally validated allosteric proteins with well-annotated allosteric sites from allosteric ligand-protein crystal complexes in Protein Data Bank (PDB) (Table S2). For these 74 allosteric proteins, we collected 624 human protein-allosteric ligand complexes from the PDB database.

Annotation of Allosteric Sites, Orthosteric Sites, and Other Sites for Human Allosteric Proteins

Allosteric Sites

We built a collection of non-redundant, high-quality benchmarking allosteric sites using 624 human allosteric complexes via the following rules: (1) only crystal complexes with allosteric ligands were included, (2) complexes bound to allosteric covalent ligands were not included, and (3) allosteric ligands were "regular" organic molecules. In addition, complexes bound to allosteric ions and peptides were not included. As a result, 501 allosteric complexes were selected. The structure coordinates for each allosteric complex were downloaded from the PDB database,[17] and the residues constituting the allosteric site were automatically extracted from the complex structure at 8 Å around the allosteric modulator site using PyMOL (The PyMOL Molecular Graphics System, v.1.7.4 Schrödinger). Then, the residues of the allosteric sites were aligned to the corresponding canonical UniProt[18] protein sequence using PDBSWS.[19] If one allosteric protein had several complexes or multiple allosteric sites, the residues from different complexes or sites were merged, resulting in a list of 74 experimentally validated allosteric proteins with well-annotated allosteric sites from the protein 3D structures.

Orthosteric Sites

We retrieved and downloaded the orthosteric complex structures for the above-mentioned 74 allosteric proteins from the PDB database[17] if they fulfilled the following two criteria: (1) the resolution of crystal structure was better than 3.0 Å and (2) the orthosteric ligands were regular small molecules. The residues constituting the orthosteric site were automatically extracted as described previously. The residues of the orthosteric sites were aligned to the corresponding canonical UniProt protein using PDBSWS.[19] Finally, we obtained a list of 48 proteins with the well-annotated orthosteric sites from protein 3D structures.

Other Sites

All cavities on the protein surface of each allosteric complex were detected and extracted by Fpocket[20] package, which can identify different types of cavities, including very small pockets, ligand binding sites, and even tunnels.[21] The "other sites" workflow with the criteria and parameters used in Fpocket is described in the following five steps. (1) All allosteric complex files in pdb format of a given protein were collected from the PDB database. (2) Cavities on each pdb file were detected and extracted into "Cavity residues" by Fpocket with the default parameters including 3 Å (−m) for the minimum radius of alpha sphere, 6 Å (−M) for the maximum radius of alpha sphere, 35 (−i) for the minimum number of alpha spheres in a pocket, 3 (−A) for the minimum number of contacting apolar atoms for an apolar sphere, 1.73 Å (−D) for the maximum distance between two alpha spheres by a Voronoi edge, 4.5 Å (−r) for the maximum cluster distance, 2 (−n) for the number of alpha spheres in a pocket that have close to alpha spheres of another pocket, 2.5 Å (−s) for the maximum distance from alpha spheres of another pocket, 0.0 (−p) for the maximum ratio of apolar alpha spheres and the number of alpha spheres in a pocket, and 2,500 (−v) for the number of iteration in Monte-Carlo algorithm. (3) Cavity residues from each pdb file of the protein were merged. (4) Residues belonging to the corresponding allosteric sites and orthosteric sites of the protein were removed from cavity residues and the

remaining residues were denoted as "other sites" of the protein. (5) The same procedure (1) to (4) of other sites above was performed on all 74 allosteric proteins. The residues of other sites were aligned to the corresponding canonical UniProt protein using PDBSWS.[19]

Collection and Preparation of Somatic Mutations

We collected and assembled somatic mutations from four resources: (1) 3,281 pairwise tumor-normal matched samples across 12 cancer types from TCGA,[4] (2) 4,938,362 mutations in 7,042 matched tumor-normal samples across 30 different cancer types/subtypes from the Sanger website,[22] (3) 1,195,223 somatic mutations in 8,207 matched tumor-normal samples across 30 cancer types/subtypes from the Elledge's Laboratory website at Harvard University,[23] and (4) the COSMIC: Catalogue of Somatic Mutations in Cancer (v.69).[24] We used ANNOVAR[25] to map these somatic mutations onto the protein sequences to identify the corresponding amino acid changes based on RefSeq ID. We calculated the functional impact score for the nonsynonymous SNVs (single-nucleotide variants) using SIFT[26] and PolyPhen-2 scores[27] via ANNOVAR. Then, we converted the RefSeq ID (accessed on September 2, 2014) to the UniProt ID (using UniProt release September, 2014) using the UniProt ID mapping tool.

Collection and Annotation of Mendelian Disease-Causing Mutations

We collected the Mendelian disease-causing mutations (missense mutations) from two resources: (1) 29,097 disease-causing mutations and 36,429 polymorphisms from the Online Mendelian Inheritance in Man (OMIM) compendium (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University) and (2) 14,444 disease-causing mutations for 574 allosteric protein-encoding genes from the Human Gene Mutation Database (professional v.2014).[28] We performed extensive informatics operations, as well as a manual curation, to combine the two data sources and remove duplicate records, resulting in a list of 12,346 disease-causing mutations and 1,980 polymorphisms in 574 allosteric protein-encoding genes.

Construction of the High-Quality Human Protein Interactome

We constructed two different yet complementary human protein interaction networks (PINs): (1) a large-scale physical PIN and (2) a kinase-substrate interaction network (KSIN). Specifically, we downloaded human physical PPIs from two resources, Protein Interaction Network Analysis (PINA, May 1, 2013) platform[29] and InnateDB,[30] to construct the physical PIN. In the KSIN, a node denotes a kinase or its substrate protein and an edge denotes a phosphorylation reaction between a kinase and its substrate protein. We collected the high-resolution kinase-substrate interaction (KSI) pairs from four databases: Phospho.ELM,[31] Human Protein Resource Database,[32] PhosphoNetworks,[33,34] and PhosphoSitePlus.[35] We implemented two data-cleaning steps. First, we defined the high-quality interactions as those that were experimentally validated in human models through a well-defined experimental protocol. Second, we annotated all protein-coding genes using the Entrez gene ID, the chromosome location, and the gene official symbols from the National Center for Biotechnology Information (NCBI) database. The detailed protocols for the construction of the PIN and KSIN are provided in our previous studies.[36,37]

Preparation of Microarray Gene Expression Data and the Co-expression Analysis

We collected microarray gene expression data across 126 normal tissues in a previous study[38] and normalized the expression values at the probe level using quantile normalization. We then computed the Pearson correlation coefficient (PCC) value using the normalized values and mapped it onto the above KSIN to build co-expressed kinase-substrate interaction network (CeKSIN), as described in two previous studies.[36,37]

Categories of Different Disease Gene Sets

Cancer-Related Genes

Here, we collected three overlapping yet complementary cancer-related gene sets, as shown below: (1) 693 significantly mutated genes (SMGs) in cancer were collected from more than 20 large-scale cancer genomic analysis projects as described in our previous study;[39] (2) 563 experimentally validated cancer genes were downloaded on February 21, 2016 from the Cancer Gene Census[26] and denoted as the CGC genes; and (3) 4,050 cancer genes were assembled in a previous study,[36] referred to here as the cancer gene atlas, namely CGA.

Other Disease Gene Sets

We collected two commonly used inherited disease gene sets: (1) 2,713 Mendelian disease genes (MDGs) were compiled from the Online Mendelian Inheritance in Man (OMIM) database[40] in December 2012 and (2) 2,123 orphan disease mutant genes (ODMGs) were collected from a previous study.[41]

Essential Genes

Essential genes, whose knockout result in lethality or infertility, are important for studying the robustness of a biological system.[42] Here, 2,719 essential genes were compiled from the OGEE database.[42]

Computing Selective Pressure and Evolutionary Rates

We calculated *dN/dS* ratios[43] to examine selective pressures on genes. Here, we used the human-mouse orthologous gene products to compute *dN* and *dS* substitution rates using the human-mouse sequence data for 16,854 gene products available in the Ensemble BioMart database. In addition, we performed an evolutionary rate ratio analysis, as described in a previous study.[44] Details of data and analyses were provided in our previous study.[36]

Inferring Protein Evolutionary Origins

Phylogenetic analysis was used to infer the evolutionary origin of a protein, referring to the approximate date that the protein originated. Here, we calculated the protein origin using ProteinHistorian.[45] Specifically, protein origin (age) was estimated by considering three factors: a species tree, a protein family database, and an ancestral family reconstruction algorithm. Furthermore, we performed an evolutionary distance analysis by comparing human sequences with orthologous sequences from other animals, as described previously.[44]

Kaplan-Meier Survival Analysis

To validate our results, we downloaded the mRNA expression profiles and the clinical data for lung adenocarcinoma[46] from TCGA website. The RNA-Seq by Expectation Maximization (RSEM) values of the mRNA[47] were used as a measure of the expression level of genes. All p values for survival analysis were calculated using the log-rank test.

Mapping of Disease-Causing Variants and Somatic Variants at the Allosteric Sites, Orthosteric Sites, and Other Sites in Allosteric Proteins

The mapping pipeline used the following steps: (1) only missense variants on the allosteric proteins with released crystal structures were kept, resulting in a list of 4,451 missense somatic variants, 2,123 disease-causing variants, and 238 polymorphisms; (2) all of the 4,451 missense somatic variants were aligned to protein sequences (using UniProt release September, 2014) using NW-align; and (3) SIFT[26] and PolyPhen-2[27] scores were calculated for each nonsynonymous somatic variant. Herein, a variant with a SIFT score < 0.05 and a PolyPhen-2 score > 0.909 was defined as deleterious (D), as described in previous studies.[26,27] Otherwise, it was defined as tolerated (T).

Description of AlloDriver

We calculated the normalized variant rate for each allosteric protein as follows:

$$\left(\frac{V_A}{P_A}\right) \Big/ \left(\frac{V_T}{P_T}\right)$$

V_A is the number of variants at the allosteric sites and V_T is the total number of variants in the corresponding protein. P_A is the number of residues at the allosteric sites and P_T is the total number of residues in the entire allosteric protein. Then, we proposed a method, named AlloDriver, to calculate the statistical significance of the variants enriched at the allosteric sites. The null hypothesis posits that somatic missense variants equally distribute at protein allosteric sites against other regions. The alternative hypothesis asserts that somatic missense variants are more likely enriched at protein allosteric sites than other sites. We performed the permutation test in AlloDriver as below:

$$P = \frac{\#\{Z_m(p) > Z_m\}}{\#\{\text{total permutations}\}}$$

A nominal P was computed for each allosteric protein by counting the number of observed missense somatic variants in a specific cancer type or pan-cancer greater than the permutations. Herein, we performed 100,000 permutations by randomly selecting the same number of at the allosteric sites on a specific protein from its total number of variants in a specific cancer type for individual analysis or in pan-cancer for pan-cancer analysis. Then, the resulting p values generated from the permutation tests were corrected as adjusted p values (q) by using Benjamini-Hochberg multiple test correction method[48] that has been implemented in the R package (v.3.1.2).[49]

Statistical Analysis

The Wilcoxon test, Kolmogorov-Smirnov tests, and Fisher's exact test were performed using the R platform (v.3.1.2).

Experimental Validation Protocols

Cell Culture

Two human lung adenocarcinoma cell lines (NCI-H23 and A549) and human embryonic kidney 293T cell line were obtained from the American Type Culture Collection (ATCC). NCI-H23 and A549 cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium, and 293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM). Cell lines were maintained in culture medium supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 100 μg/mL streptomycin at 37°C in a humidified atmosphere containing 5% CO_2.

Plasmid Construction

The human PDE10A2 expression construct pCMV-PDE10A2 (GenBank: NM_006661.3) was purchased from Biogot Technology. Then, the full-length and mutant (QuikChange Site-Directed Mutagenesis Kit from Agilent) cDNAs were amplified using the 2 × Pfu (PCR) Master Mix (Lifefeng) and sub-cloned into the XbaI and BamHI sites on a lentiviral vector pCDH-CMV-MCS-EF1-copGFP (System Biosciences). All plasmids were verified by sequencing. The gene sequence used for the construction of pCMV-PDE10A2 is provided in Table S3.

Production of the Lentivirus and the Infection of NCI-H23

293T cells in 10 cm diameter dishes were transfected with a combination of the expression vectors of the human wild-type or mutant PDE10A2 and the lentiviral packaging vectors psPAX2 (addgene, plasmid #12260) and pMD2.G (addgene, plasmid #12259) using the X-tremeGENE 9 DNA Transfection Reagent (Roche). The supernatant of the cultured cells was replaced with fresh medium 4–6 hr after transfection. After incubation for 48–72 hr, the supernatants of the transfected cells containing viruses were harvested and filtered through a 0.45 μm syringe filter, and the viruses were used to infect the NCI-H23 cells immediately or frozen at −80°C. If required, viruses were concentrated by ultracentrifugation at 28,000 rpm for 2 hr at 4°C. The pellets were resuspended in PBS containing 2% FBS and aliquoted for storage at −80°C. NCI-H23 cells were seeded into 6 cm diameter dishes and infected with the concentrated lentivirus the next day, and the polybrene with a final concentration of 8 μg/mL was added to the infected cells to enhance the infection efficiency. To obtain higher infection efficiencies, the infected NCI-H23 cells were sorted using a flow cytometry sorter (Beckman). Finally, the GFP-positive rate of these stable NCI-H23 cell lines was found to be greater than 95%.

Reagents and Antibodies

Two compounds, PF-2545920 and dipyridamole, were purchased from Selleckchem. The antibody of PDE10A2 (88 KD) was purchased from Abcam. The corresponding secondary antibody and ACTIN-HRP were purchased from Cell Signaling Technology.

Western Blotting Analysis

Cells were lysed in 2× SDS lysis buffer (100 mM Tris HCl [pH 6.8], 200 mM DTT, 4% SDS, 0.2% bromophenol blue, and 20% glycerin). Proteins in the samples were separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred by electroblotting onto PVDF membranes (Millipore). The PVDF membranes were blocked with 5% non-fat milk at room temperature for 1 hr and then incubated with the appropriate primary antibody at 4°C overnight. After additional TBST washes, the membranes were incubated with the corresponding horseradish peroxidase-conjugated secondary antibodies for 1 hr at room temperature and detected using the enhanced chemiluminescence method (Millipore).

Cell Growth Detection

Cells were seeded in 96-well plates at a density of 4,000 cells and incubated for 24 hr. Then, the cells were treated with the specified compound or the vehicle control and incubated for another 72 hr. The inhibition of cell growth caused by the treatments was determined using the CellTiter 96 Aqueous One Solution Cell Proliferation Assay (MTS) (Promega). To validate the impact of the mutant on lung adenocarcinoma cancer cell growth, the stable cell line NCI-H23 and corresponding control cells were seeded into 96-well plates at a density of 600 cells cultured in medium containing

1% FBS to minimize the interference of serum. Then, cell growth was detected at the indicated time using the CellTiter 96 Aqueous One Solution Cell Proliferation Assay (MTS). The assays were conducted according to the manufacturer's instructions, and the absorbance value (optical density) of each well was measured at 490 nm using a microplate reader. The absorbance at 630 nm was subtracted from this number, as the basic value. All experiments were performed at least three times.

Colony Formation Assay

Cells were plated in 6-well culture plates at 600 cells/well, and each cell group had three wells. After incubation for another 12 days at 37°C, the cells were washed twice with cold phosphate buffer saline, fixed using ice-cold 100% methanol, and stained with a 0.5% crystal violet solution. Then, the stained colonies were washed with double-distilled water and photographed. The number of the colonies containing >50 cells was counted. All assays were independently performed in triplicate.

Experimental Design and Data Analysis

With regard to the effects of the drugs on cell growth, the IC_{50} values, which were the concentrations of the compounds when cell viability was 50%, were determined. All experiments were repeated a minimum of three times to determine the reproducibility of the results. All error bars represent the SEM. Statistical analysis was performed using Student's t test. A p value < 0.05 was considered to be statistically significant.

Results

An Integrative Genomic Workflow to Elucidate Cancer-Associated Protein Allosteric Dysregulation

We constructed a global human allosteric protein catalog based on the Allosteric Database (ASD) developed by our group.[16] We carefully curated records from the ASD to produce a high-quality allosteric protein catalog. It included 574 human gene products (proteins), in which 74 proteins have experimentally validated allosteric sites according to their allosteric ligand-protein complex structures in the PDB database (see Material and Methods). The functional classes for the human allosteric proteins annotated from UniProt[18] are shown in Figure 1A. The most abundant allosteric proteins were transferases (21%) and hydrolases (17%). Then, we collected somatic mutations from TCGA, the Catalogue of Somatic Mutations in Cancer (COSMIC) database, and other public domains (see Material and Methods). In total, we obtained 47,364 somatic missense mutations from 6,958 pairwise tumor-normal matched pairs across 33 cancer types on 574 allosteric protein-coding genes. Figure 1B shows the somatic missense variant load for the allosteric proteins across 12 common cancer types or subtypes with unique sample IDs. To further explore the relationship between these variants and their associated cancer types, we designed a pipeline to annotate the variants at the allosteric sites, orthosteric sites, and other sites in the allosteric proteins (Figure 1C, see Material and Methods). Next, we developed a statistical model to identify the functional somatic variants that allosterically alter protein activity in pan-cancer as well as each individual cancer type (Figure 1D). Finally, we tested our model predictions both computationally (Figure 1E) and experimentally (Figure 1F) using NSCLC as a case study.

Network Characteristics of Allosteric Proteins in the Human Protein Interaction Network

To examine the biological functions of the allosteric protein catalog, we investigated the topological network features (e.g., the connectivity) of allosteric proteins in the human PIN. Considering that the current publicly available human PIN has data bias and is incomplete, well-studied human kinome data may provide more valuable features by local ecosystem. We constructed two complementary human PINs, a global physical PIN and a KSIN, based on our two previous studies (see Material and Methods).[36,37] Figure 2A shows that the connectivity of allosteric proteins is significantly stronger than that of non-allosteric proteins in both the KSIN ($p = 1.9 \times 10^{-10}$, Wilcoxon test) and the physical PIN ($p = 6.3 \times 10^{-43}$). A previous study has suggested that the kinome network plays important biological roles in cancer and 16% of the allosteric proteins are well-known kinases (Figure 1A).[36] To further examine the functional roles of allosteric proteins at the network level, we examined the gene co-expression distribution for the allosteric protein-protein pairs using the human kinome data.[36] We calculated the PCC for the gene-gene pairs using microarray gene expression data from 126 normal tissues, as described in our previous study.[36] We mapped the PCC value onto the KSIN to build a CeKSIN. Here, we defined an allosteric kinase-substrate interaction (KSI) pair as either one or two proteins in a pair that is/are allosteric protein(s) in CeKSIN. Figure 2E indicates that the allosteric CeKSI pairs are more likely to be the highly co-expressed KSI pairs ($p = 1.3 \times 10^{-4}$, Fisher's exact test). Thus, allosteric protein-coding genes tend to be highly co-expressed in CeKSIN, suggesting their critical biological roles in network perturbations.

Evolutionary Trajectories of Allosteric Proteins

We further examined the selective pressure and evolutionary rates of allosteric proteins. We calculated the non-synonymous and synonymous substitution rate ratio (the *dN/dS* ratio) using human-mouse orthologous gene products (see Material and Methods). A *dN/dS* ratio of 1 signifies neutral evolution, whereas a ratio < 1 indicates purifying selection, and a ratio > 1 indicates positive Darwinian selection. Figures 2C and 2D show that allosteric proteins have a lower *dN/dS* ratio and a lower evolutionary rate ratio than non-allosteric proteins, suggesting that allosteric proteins tend to undergo strong purifying selection (meaning that the *dN/dS* ratio is < 0.1). The evolutionary history of a protein sequence often reflects its functionally evolutionary trajectory. Next, we examined the evolutionary origin of allosteric proteins. Here, phylogenetic analysis was used to infer the evolutionary origin of a protein, referring to the approximate date that the protein originated. Specifically, we calculated the protein origin by considering three factors: a species tree, a protein family database,

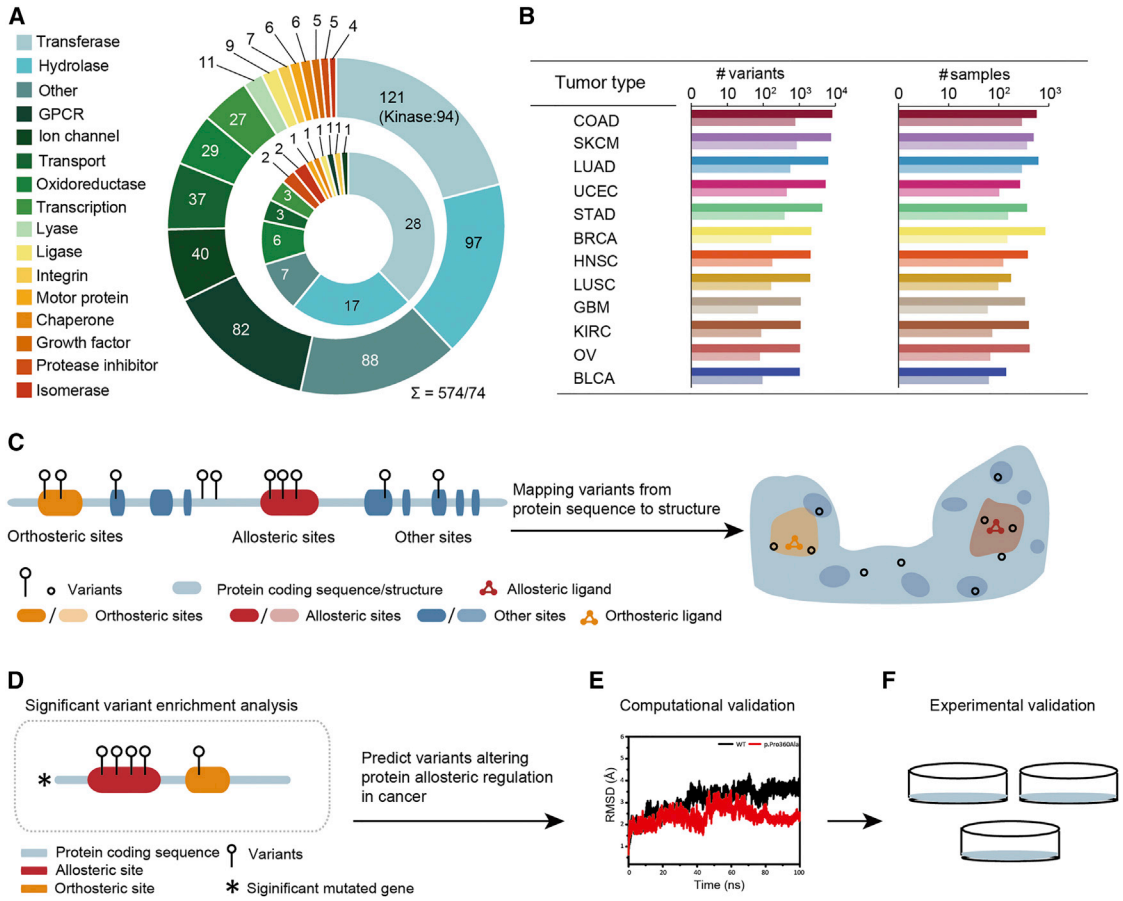

Figure 1. Integrative Genomic Workflow and Data Analysis

(A) The functional groups of the allosteric proteins. There are 574 human allosteric proteins from the allosteric database (ASD), each of which has been experimentally validated by at least three cases of evidences in crystal structure complex or biochemistry (such as site-directed mutagenesis, cooperativity of kinetic effect from two ligands, and uncompetitive binding assay with chromatography, etc.). Only 74 of these 574 proteins demonstrated the exact location of allosteric sites by the crystal structures of allosteric ligand-protein complex in the PDB database.

(B) The number of variants in the allosteric proteins and corresponding samples in 12 common cancer types Abbreviations are as follows: COAD, colon adenocarcinoma; SKCM, skin cutaneous melanoma; LUAD, lung adenocarcinoma; UCEC, uterine corpus endometrial carcinoma; STAD, stomach adenocarcinoma; BRCA, breast invasive carcinoma; HNSC, head and neck squamous cell carcinoma; LUSC, lung squamous cell carcinoma; GBM, glioblastoma multiforme; KIRC, kidney renal clear cell carcinoma; OV, ovarian serous cystadenocarcinoma; and BLCA, bladder urothelial carcinoma. For each cancer type, the upper bar denotes the number of variants or samples in the 574 allosteric proteins and the bottom bar denotes the number of variants or samples in the 74 allosteric proteins with released allosteric sites.

(C) The mapping of the mutations detected in the DNA sequences onto orthosteric, allosteric, or other sites in allosteric proteins.

(D) The identification of significantly mutated allosteric proteins in cancer.

(E) The computational validation of the predicted cancer-associated variants.

(F) The experimental validation of the predicted cancer-associated genes as well as the variants.

and an ancestral family reconstruction algorithm, using ProteinHistorian.[45] Figure 2B shows that most of the allosteric proteins had a divergence time from 910 million years ago (Mya) to 4,200 Mya with average time of divergence at 1,648.6 Mya, which was significantly older than that of non-allosteric proteins (1,178.4 Mya, $p = 2.1 \times 10^{-8}$, Kolmogorov-Smirnov test). Interestingly, a threshold age of 1,600 Mya happens to be the transition time between Paleoproterozoic and Mesoproterozic, when aerobic respiration began to emerge.[36,50] Thus, our evolutionary trajectories analysis suggested that the transition between anaerobic respiration and aerobic respiration may have led to the emergence of protein allostery. Altogether, these results suggest that allosteric regulation is functionally inherent and under strong purifying selection during protein sequence evolution.

Perturbations of Allosteric Proteins at Allosteric Sites Reflects Disease Etiology

The aforementioned network topology and protein evolutionary analyses indicated a critical biological role for allosteric proteins. The dysfunction of allosteric proteins may have important effects on mediating human diseases such as cancer. To investigate their effect, we performed a

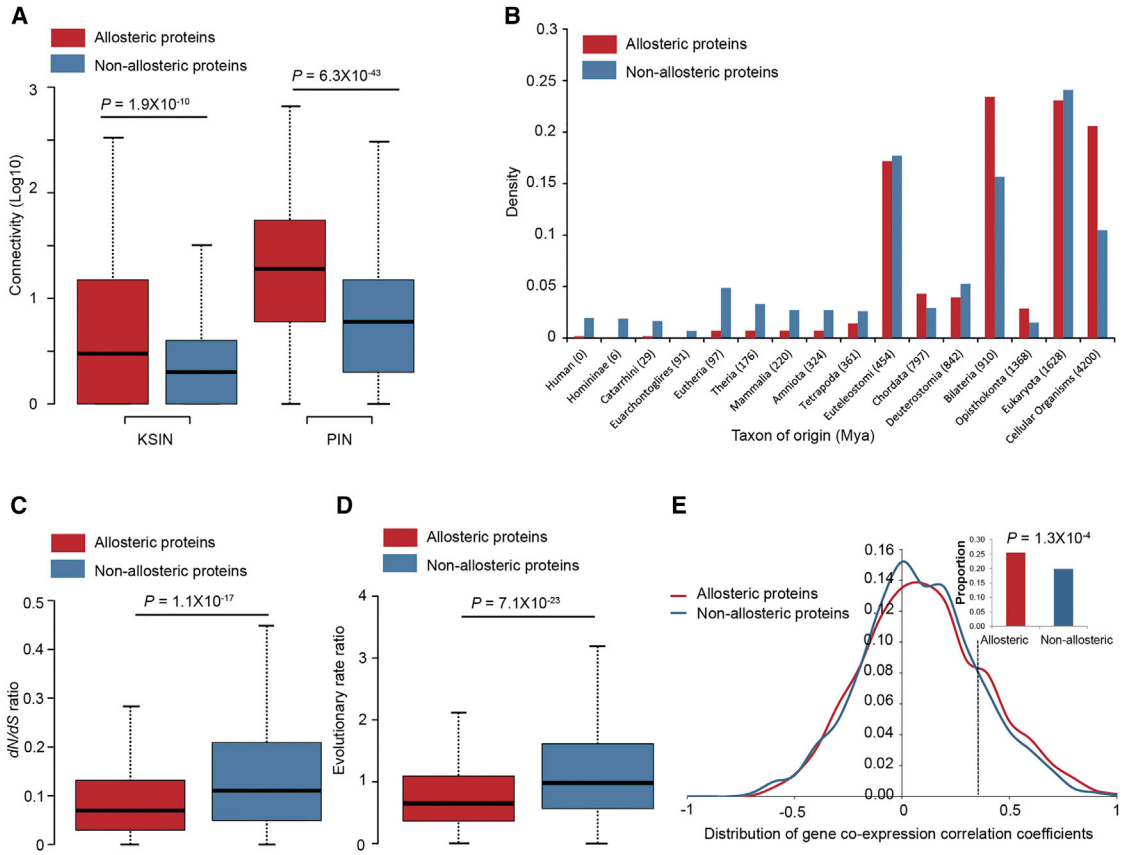

Figure 2. Network Characteristics and Evolutionary Trajectories of Allosteric Proteins
(A) The connectivity distribution for allosteric proteins versus non-allosteric proteins in the kinase-substrate interaction network (KSIN) and the protein interaction network (PIN).
(B) Distribution of taxon of origin (million years ago [Mya]) for allosteric proteins versus non-allosteric proteins.
(C) Distribution of the *dN/dS* ratio for allosteric proteins versus non-allosteric proteins.
(D) Evolutionary rate ratio for allosteric proteins versus non-allosteric proteins.
(E) Distribution of gene co-expression correlation coefficients for allosteric proteins versus non-allosteric proteins in the co-expressed kinase-substrate interaction network. Allosteric proteins (red) denote either protein that is allosteric protein in a pair. Non-allosteric proteins (blue) denote two proteins that are non-allosteric proteins in a pair.
p values in (A), (C), and (D) were calculated via Wilcoxon rank-sum test. p values in (E) were calculated by Fisher's exact test.

diseasome analysis focusing on allosteric proteins. Among the genes encoding 574 allosteric proteins (in short allosteric genes below), we found that 340 allosteric genes were known cancer-associated genes ($p = 3.0 \times 10^{-89}$, Fisher's test, including 76 cancer driver genes [$p = 4.2 \times 10^{-24}$]) and that 230 genes were Mendelian or orphan disease-causing genes ($p = 4.9 \times 10^{-53}$). Hence, allosteric regulations are significantly involved in cancer (Figures 3A and 3B). Meanwhile, 47,364 somatic missense mutations, 12,346 disease-causing mutations, and 1,980 polymorphisms were collected for the 574 allosteric genes (see Material and Methods), and 4,451 somatic missense variants, 2,123 disease-causing variants, and 238 polymorphisms were mapped onto 74 allosteric protein sites in the released 3D structures. First, we classified the 4,451 somatic variants into two categories: 1,990 deleterious variants (SIFT scores < 0.05 and PolyPhen-2 score > 0.909) and 2,461 tolerated variants. In addition, we further classified all potential sites throughout the protein structures into three groups: allosteric sites, orthosteric sites, and other sites (see Material and Methods). Figure 3C revealed that deleterious variants (683 of 1,990) were significantly enriched in allosteric sites ($p = 4.2 \times 10^{-4}$, Wilcoxon test) compared to tolerated variants (252 of 2,461). The high enrichment of deleterious variants was also found at orthosteric sites ($p = 2.0 \times 10^{-5}$, Figure 3C), whereas there was no significant difference between deleterious and tolerated variants at the other sites ($p = 0.76$, Figure 3C). In addition, we found a similar trend for disease-causing variants: they were significantly enriched at allosteric sites ($p = 1.2 \times 10^{-6}$, Wilcoxon test, Figure 3D) and orthosteric sites ($p = 1.5 \times 10^{-5}$, Figure 3D) compared with polymorphisms, while not at the other sites ($p = 0.48$, Figure 3D). Interestingly, we did not observe the statistical difference for the distribution of both deleterious somatic missense variants ($p = 0.2441$) and disease-causing variants ($p = 0.5176$) at the allosteric sites from that at the orthosteric sites. Altogether, allosteric sites seem to

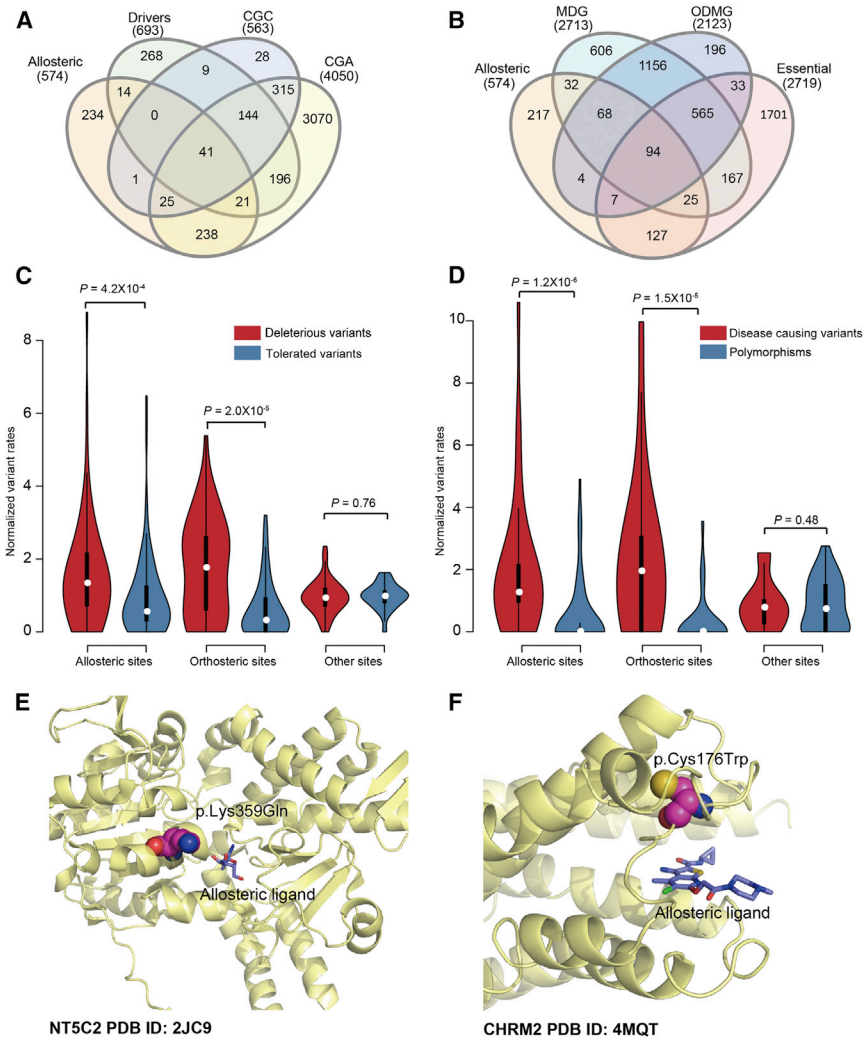

Figure 3. Perturbations of the Signaling Network of Allosteric Proteins Reflect Disease Etiology

(A) Overlaps among allosteric proteins, cancer driver genes (Drivers), the Cancer Gene Census (CGC), and the Cancer Gene Atlas (CGA) were shown by Venn diagram.

(B) Overlaps among allosteric proteins, Mendelian disease genes (MDG), orphan disease mutant genes (ODMG), and essential genes (Essential) were shown by Venn diagram.

(C) Distributions of deleterious somatic missense variants versus tolerated variants were depicted by Bean plots at allosteric sites, orthosteric sites, and other sites.

(D) Distributions of disease-causing variants versus polymorphisms were depicted by Bean plots in allosteric sites, orthosteric sites, and other sites, respectively.

(E) p.Lys359Gln located at the allosteric site of NT5C2 mediates chemotherapy resistance in relapsed acute lymphoblastic leukemia (ALL).

(F) The disease-causing p.Cys176Trp variant in individuals with dilated cardiomyopathy was located at the allosteric site of CHRM2.

p values in (C) and (D) were calculated via Wilcoxon test.

produce crucial effects in protein function in mediating disease pathology, like at orthosteric sites.[51] Figures 3E and 3F illustrate two examples of allosteric proteins harboring somatic or disease-causing variants at their allosteric sites. Specifically, p.Lys359Gln at the allosteric site of NT5C2 is an activating substitution that mediates chemotherapy resistance in relapsed acute lymphoblastic leukemia (ALL), resulting in an increased NT5C2 activity by mimicking the effect of positive allosteric regulators (Figure 3E).[52] A p.Cys176Trp alteration, located at the allosteric site of the M2-muscarinic acetylcholine receptor (CHRM2), has been identified to be a disease-causing variant in individuals with dilated cardiomyopathy (MIM: 115200) (Figure 3F).[53,54] Collectively, these observations suggest that allosteric proteins altered by somatic variants or disease-causing variants at allosteric sites may perform indispensable roles in human diseases.

Allosteric Regulation Altered by Somatic Variants in Pan-cancer

We next developed a statistical approach, named AlloDriver, to prioritize significantly mutated allosteric proteins in cancer. The hypothesis of our statistical model was that a protein that harbors the enriched somatic missense variants at its allosteric sites is more likely to be involved in cancer (Figure 3C). We performed a pan-cancer analysis based on 47,364 somatic missense variants collected from 6,958 tumor-normal matched samples on 74 allosteric proteins with the experimentally validated allosteric sites. We identified three proteins harboring significantly enriched missense variants at their allosteric sites during our pan-cancer analysis ($q < 0.1$) (Figure 4A). These three proteins were BRAF[55] ($q < 10^{-6}$), HRAS[56] ($q = 0.023$), and AKT1[57,58] ($q = 0.048$), which are well-known cancer-associated proteins. To explore more allosteric proteins altered by somatic variants, we next examined the allosteric proteins with p values < 0.05 as well as at least two variants at allosteric sites. In addition, we found six potential proteins: SERPINC1 ($p = 0.013$), CHRM2 ($p = 0.016$), GCK ($p = 0.020$), MAPK8 ($p = 0.021$), LTA4H ($p = 0.030$), and AR ($p = 0.045$), as shown in Figure 4A. Among these six proteins, four (CHRM2,[59] MAPK8,[60] LTA4H,[61] and AR[62]) have been reported to be involved in tumorigenesis and tumor progression in vitro and in vivo. Remarkably, two original proteins, SERPINC1 and GCK, were predicted to be significant by our pan-cancer model, which could help to uncover the new functional roles of

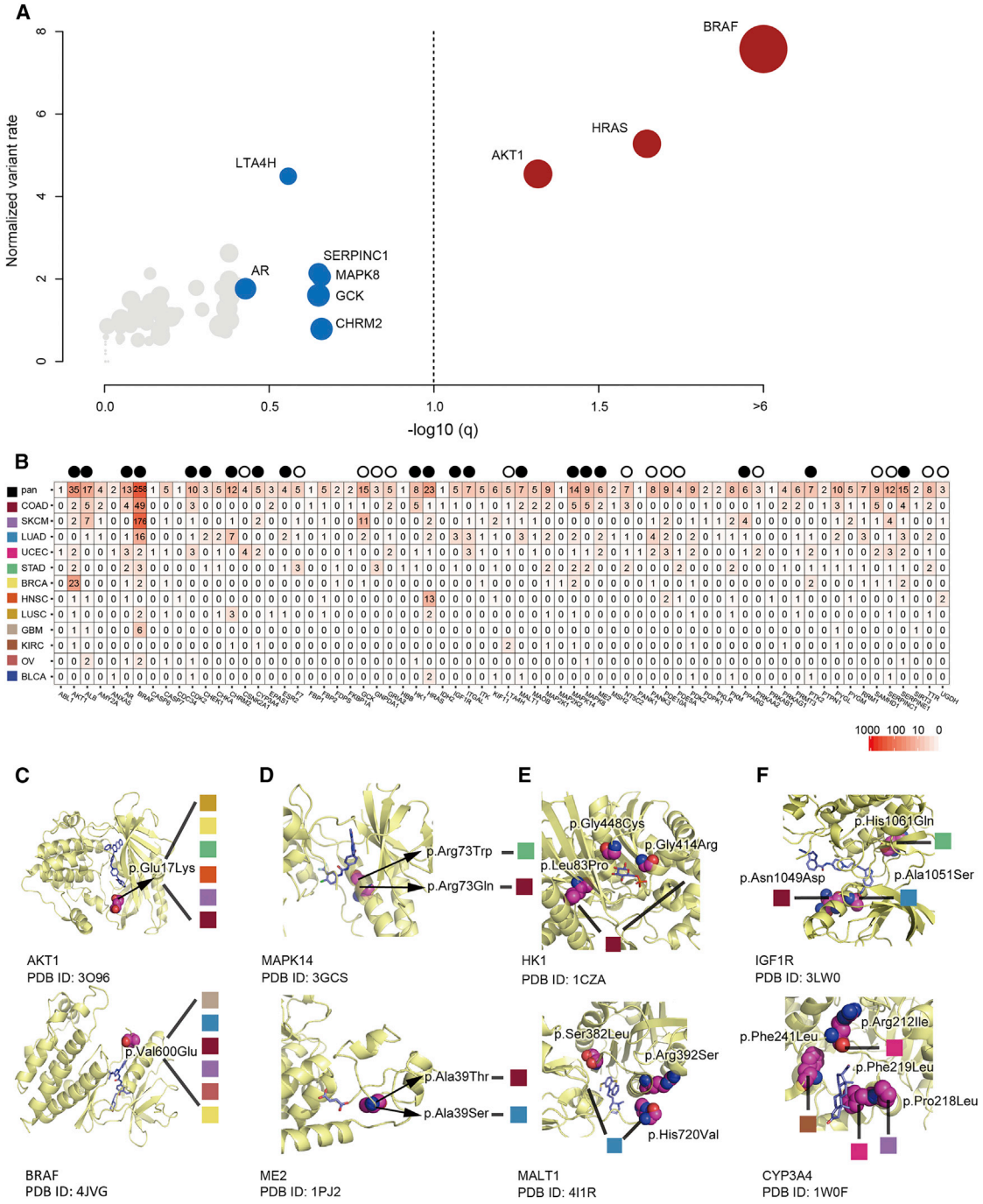

Figure 4. Discovery of Mutated Allosteric Proteins in Pan-cancer and Individual Cancer Analyses

(A) Pan-cancer analysis of allosteric proteins using somatic missense mutations observed in approximately 7,000 matched tumor-normal samples. The x axis is $-\log_{10}(q)$ of variants and the y axis is the normalized variant rates at 74 allosteric proteins with experimentally validated allosteric sites according to the crystal structures of allosteric ligand-protein complex. The gray dashed line represented the adjusted p value (q) of 0.1. The size of the dots is proportional to the $\log_{10}$ (# of variants) value. Proteins with $q < 0.1$ are colored in red, proteins with $p < 0.05$ are colored in blue, and the others are colored in gray.

(B) A heatmap depicting the frequency of variants at the allosteric sites of 74 allosteric proteins in each cancer type and pan-cancer. The mutated allosteric proteins predicted by our approach in the pan-cancer and individual cancer analyses are marked with circles (mutated proteins validated by previous literatures are colored in black and unreported ones are colored in white).

(C–F) Eight representatively mutated proteins harboring the enriched somatic missense variants at their allosteric sites. The squares in color represent individual cancer types the same as those in (B).

the two proteins in cancer. Thus, our pan-cancer analysis suggests that protein allosteric dysregulation could be a key factor during tumorigenesis and provides a potential strategy to identify mutated proteins with the perturbed allosteric regulations caused by somatic variants.

Identification of Potential Mutated Allosteric Proteins in 12 Individual Cancer Types

We further investigated the mutated allosteric proteins that harbor enriched somatic missense variants at their allosteric sites for individual cancer types/subtypes using AlloDriver. As a result, we observed 35 mutated allosteric proteins across 12 cancer types (Table S4). The predicted mutated allosteric proteins are marked with a circle in Figure 4B and the frequency and the number of variants at allosteric sites for the allosteric proteins in each individual cancer type and pan-cancer are also shown. Of 35 predicted allosteric proteins, 20 proteins have been shown to be associated with the initiation and progression of specific individual cancer in previous reports (black circle in Figure 4B, Table S4). For example, somatic variants in AKT1 ($q < 0.005$), BRAF ($q < 0.005$), HRAS ($q < 0.005$), PTK2 ($p < 0.05$), and AR ($p < 0.05$) were reported to significantly alter protein allosteric regulation in multiple cancer types, including colon (COAD), skin (SKCM), lung (LUAD and LUSC), uterine (UCEC), stomach (STAD), breast (BRCA), head and neck (HNSC), glioblastoma (GBM), ovarian (OV), and bladder (BLCA). Figure 4C shows the structure location of two well-known driver variants at allosteric sites that induces allosteric dysregulation in multiple cancer types, p.Glu17Lys on AKT1 and p.Val600Glu on BRAF. Although molecular mechanism remains to be studied thoroughly, MAPK8 ($p = 0.0034$) and HK1 ($p = 0.0062$) in COAD, PPARG ($p = 0.0090$) in SKCM, CHRM2 ($p = 0.0002$), MALT1 ($p = 0.0073$), IGF1R ($p = 0.0110$), ESR2 ($p = 0.0171$), ME2 ($p = 0.0198$), CHEK1 ($p = 0.0243$), and ITGAL ($p = 0.0261$) in LUAD, CYP3A4 ($p = 0.0034$) and CDK2 ($q = 0.0905$) in UCEC, SERPINE1 ($p = 0.0138$) and MAPK14 ($p = 0.0276$) in BRCA, and ALB ($p = 0.0083$) in OV were reported to contribute to tumorigenesis (Table S4), which is in good agreement with our prediction according to the model. More interestingly, the evidences of connection between somatic variants at allosteric sites and special cancer phenotype suggest that pathological mechanism of these mutated proteins could derive from perturbed allosteric regulation.

We further explored the relationship between structure and function of variants at an allosteric site in cancer. More than 500 crystal structures of the allosteric proteins were selected from the PDB database and 4,451 missense somatic variants at 74 allosteric sites from 12 cancer types were mapped into the structures in our analysis. We observed four classical variant patterns at an allosteric site (Figure S1): (1) the same variant on the same residue contributes to multiple cancer types, e.g., AKT1 and BRAF (Figure 4C); (2) the different variants on the same residue contribute to different cancer types, e.g., MAPK14 and ME2 (Figure 4D); (3) the variants on the different residues contribute to the same cancer type, e.g., HK1 and MALT1 (Figure 4E); and (4) the variants on the different residues contribute to different cancer types, e.g., IGF1R and CYP3A4 (Figure 4F). These comprehensive patterns suggest that allosteric regulation of variant from allosteric site to orthosteric site in a protein may be highly dependent of the protein's complex partners and network in various cancer types.

Besides the 20 validated proteins mentioned above, we found that 15 potential proteins were significantly mutated in special individual cancer type (white circle in Figure 4B, Table S4) such as PDE10A, GCK, SERPINC1, etc. Among 15 proteins, PDE10A was predicted to enrich missense somatic variants at allosteric site for up to three individual cancer types, ranging from UCEC ($p = 0.00378$), HNSC ($p = 0.007$), and LUAD ($p = 0.040$) (Figures 5A and 5B). Hence, we selected PDE10A as a candidate to experimentally examine its functional role in lung cancer as a case study.

Pharmacological Inhibition of PDE10A Suppresses Growth of Lung Cancer Cells

Cyclic nucleotide phosphodiesterases (PDEs) catalyze the degradation of the important second messengers, namely cyclic nucleotides cAMP and cGMP,[63] and cAMP is able to allosterically stimulate the catalysis of PDE10A by binding to an allosteric site in the GAF domain of PDE10A.[64] Herein, we found that PDE10A may be involved in lung cancer by altering allosteric regulation (Figure 5B). To examine the clinical features of *PDE10A* in lung cancer, we correlated the expression of *PDE10A* with the overall survival of LUAD-affected individuals from TCGA.[46] The Kaplan-Meier survival analysis (see Material and Methods) revealed that high *PDE10A* expression was significantly correlated with poor prognosis in LUAD-affected individuals ($p = 0.03$, Figure 5C). Figure 5D shows the elevated expression of PDE10A in two NSCLC cell lines represented for LUAD: NCI-H23 and A549. Remarkably, the pharmacological inhibition of PDE10A by both a known PDE10A selective inhibitor (PF-2545920) and a phosphodiesterase inhibitor (dipyridamole) showed potential anti-proliferative effects, with IC_{50} values of 13.5 μM and 33.9 μM, respectively (Figures 5E and 5F). Thus, PDE10A may play a potential role in LUAD, and these known PDE10A inhibitors may provide a potential pharmacological strategy for the targeted therapy in lung cancer.

Experimental Validation of a Potential Oncogenic Role of p.Pro360Ala on PDE10A in NSCLC Cells

Among the 141 reported missense variants found in PDE10A, p.Pro360Ala was identified as a deleterious variant (SIFT = 0.04 and PolyPhen-2 = 0.998) in LUAD. To investigate the functional role of p.Pro360Ala (Figure 6A), we first performed a molecular dynamic (MD) simulation for the wild-type (WT) PDE10A versus the mutated (p.Pro360Ala) PDE10A (Figure S2). For the two systems, the time dependence of the root-mean-square deviation (RMSD) of the backbone atoms relative to the initial structure and the root mean-square fluctuation (RMSF) were calculated along the simulation trajectories. Figure 6B revealed that the RMSD of PDE10A with

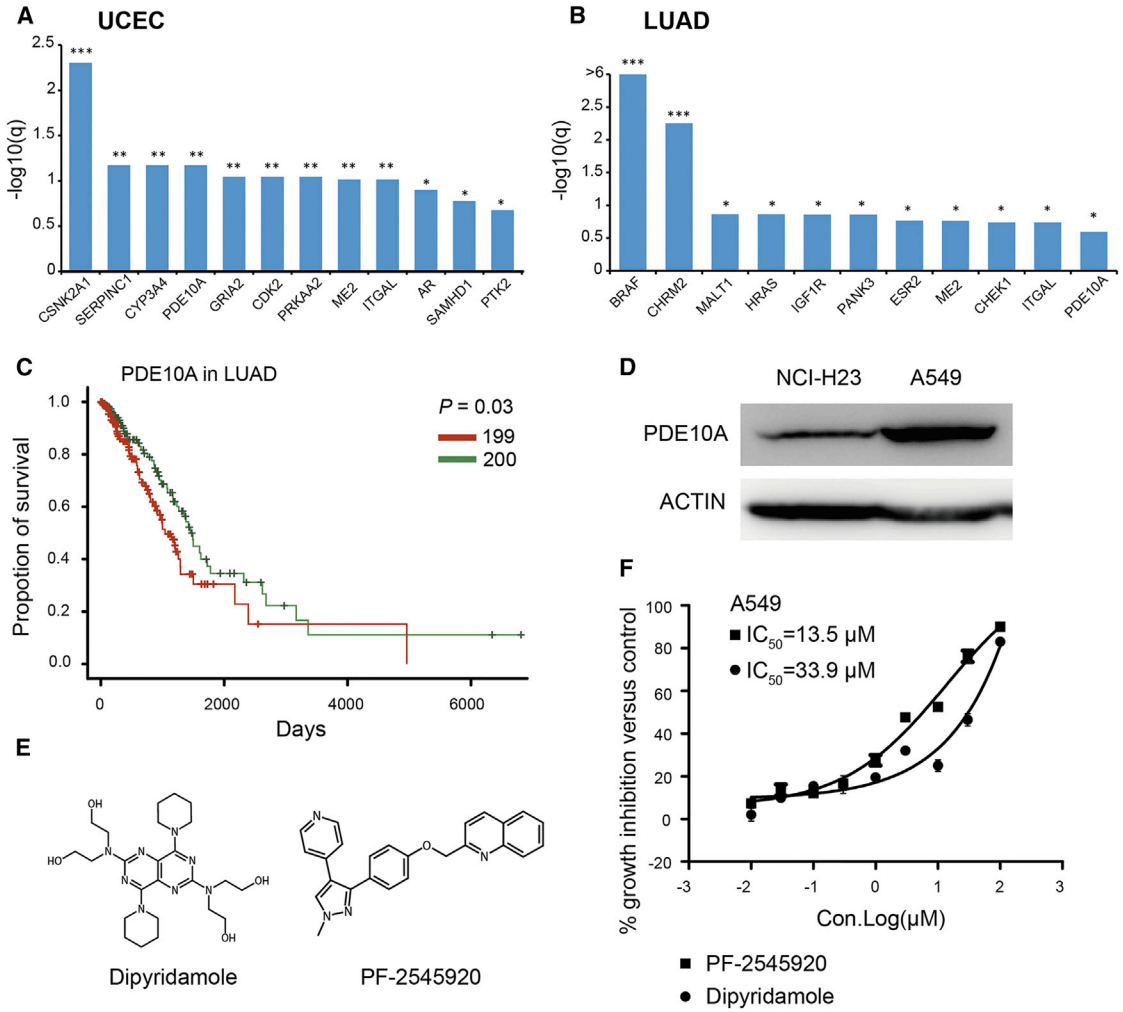

Figure 5. Experimental Validation of Functional Role to PDE10A in Lung Adenocarcinoma
(A) The top proteins of the individual analysis in uterine cancer (UCEC). ***q < 0.05, **q < 0.1, *p < 0.05.
(B) The top proteins of the individual analysis in lung adenocarcinoma (LUAD). ***q < 0.05, **q < 0.1, *p < 0.05.
(C) The Kaplan-Meier survival curves for *PDE10A* in LUAD. Individuals were separated into the high (red) and low (green) expression groups, as measured by the median gene expression level (RNA-seq). The p value in survival analysis was performed using a log-rank test.
(D) The PDE10A protein expression level in two human lung adenocarcinoma cell lines, NCI-H23 and A549, as determined by western blotting.
(E) The chemical structures of two PDE10A inhibitors, dipyridamole and PF-2545920.
(F) Cell viability assays for dipyridamole and PF-2545920 using A549 cells. IC_{50} represents half maximal inhibitory concentration.
All error bars represent the SEM from three to six independent experiments.

p.Pro360Ala was more stable than that of WT PDE10A, suggesting a positive effect on the maintenance of the PDE10A conformation by p.Pro360Ala. Meanwhile, the RMSF profile of PDE10A with p.Pro360Ala showed lower atomic fluctuations at residues 190–260 and residues 280–320 (part of allosteric site, Figure S2). These results suggest that at the allosteric site of PDE10A, p.Pro360Ala stabilized the conformation of the entire protein by reducing the flexibility of the key residues. In addition to the conformational evidence, the energy landscapes of WT versus p.Pro360Ala PDE10A were calculated and compared using principal-component analysis (PCA) profiles (Figure 6C). In WT PDE10A, there were at least two distinct energy wells in its conformational ensemble, and the active conformation from the crystal structure was found to be unfavorable in terms of energy. For PDE10A with p.Pro360Ala, there was only one energy deep well, and the active conformation became an energy-favorable state. Therefore, the computational simulations suggested that p.Pro360Ala located at allosteric site may stabilize the active conformation of PDE10A and retain a favorable energy state, leading to its persistent activation in the pathogenesis of LUAD.

Next, we experimentally tested the functional roles of p.Pro360Ala in LUAD. Figure 6D shows the elevated expression of PDE10A with p.Pro360Ala in NCI-H23 cell lines compared with vector control cells. We observed a 67% increase in viable cell number in cells that stably overexpressed PDE10A with p.Pro360Ala compared with vector control cells over 5 days (Figure 6E). In addition, in the

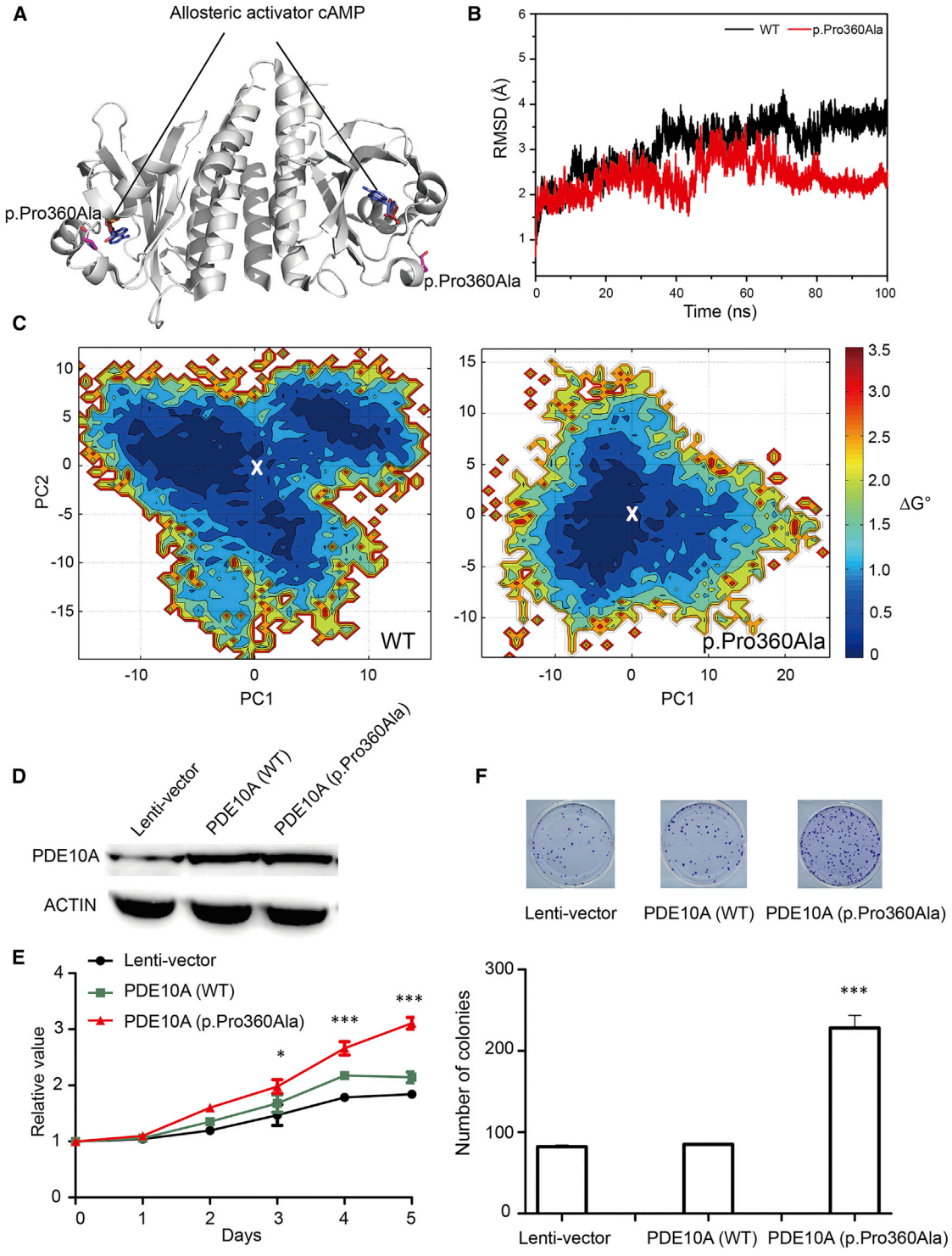

Figure 6. The p.Pro360Ala in PDE10A Is Potentially Oncogenic in Lung Adenocarcinoma
(A) The location of p.Pro360Ala in three-dimensional crystal structure of PDE10A (PDB: 2ZMF).
(B) The distribution of the root-mean-square displacement (RMSD) values of the backbone atoms between the wild-type complex and p.Pro360Ala mutant complex.
(C) Principal-component analysis (PCA) of the conformational changes on wild-type (left) and p.Pro360Ala mutant (right) PDE10A. Symbol "X" in white represents the closest conformation compared to the crystal structure of PDE10A.
(D) The overexpression of PDE10A or PDE10A p.Pro360Ala in NCI-H23 cells.
(E) The relative growth curve of PDE10A or PDE10A p.Pro360Ala in the cells that overexpress each protein in low serum medium (1% FBS, 600 cells/well). *$p < 0.05$; ***$p < 0.001$.
(F) Colony formation assays on both wild-type and p.Pro360Ala mutant PDE10A in NCI-H23 cells.
All error bars represent the SEM from three to six independent experiments.

colony-formation assay, a significant increase in colony numbers was observed in cells stably overexpressing PDE10A with p.Pro360Ala compared with WT cells ($p < 0.001$, Figure 6F). Overall, our preliminary experimental data indicated that p.Pro360Ala on PDE10A would be potentially oncogenic in LUAD. Further study is warranted to determine the roles of p.Pro360Ala on PDE10A in multiple lung cancer cell lines and in vivo.

Discussion

The allosteric regulation is an intrinsic function of protein under many physiological and pathological conditions, including cancer. However, there is lack of systematic investigation of protein allosteric regulation perturbations caused by somatic mutations in cancer. In this study, we performed comprehensive analyses to explore the dysregulation of allosteric protein function altered by somatic mutations in approximately 7,000 cancer genomes across 33 cancer types. We found that allosteric proteins tended to have stronger connectivity in the constructed human PIN and KSIN, with high selectivity pressure, and in their ancient evolutionary histories. Specifically, allosteric proteins are more likely to be highly co-expressed in the gene co-expressed KSIN, suggesting their critical roles in mediating cellular function. In addition, we showed that somatic deleterious variants and germline disease-causing variants were significantly enriched for protein allosteric sites compared with tolerated ones and polymorphisms, further suggesting the important biological role of allosteric regulation in the etiology of human diseases such as cancer.

Several previous studies have suggested that somatic missense variants often change protein functional regions on protein three-dimensional structures, such as ligand-protein binding sites[7,65] and protein-protein interfaces.[66] Our observations on protein allosteric dysregulation by somatic variants (Figure 3) are consistent with those previous studies.[7,65,66] In addition, we further developed a permutation statistical model AlloDriver to focus on identifying disease-associated cancer mutated allosteric proteins at particular function regions, allosteric sites, when analyzing more than 47,000 somatic missense mutations. We identified a series of mutated allosteric proteins that harbor enriched somatic variants at their allosteric sites during our pan-cancer and individual cancer-type analyses. Several well-known cancer gene-encoding proteins, such as BRAF, HRAS, and AKT1, often harbor somatic hotspots at their allosteric sites. In addition, we also found allosteric regulation-specific variants and 15 potential mutated proteins with altered allosteric function in multiple cancer types. Taken together, this study systematically examines allosteric perturbations caused by somatic mutations in large-scale cancer genomes, and we not only detected mutated proteins for further experimental investigation but also facilitated the understanding of important original biological consequences for somatic mutations mediating tumor initiation and progression.

More importantly, we experimentally validated that PDE10A may mediate NSCLC cell growth. In addition, high expression of PDE10A is significantly associated with poor survival in LUAD-affected individuals.[46] Moreover, the pharmacological inhibition of PDE10A by existing PDE10A small molecule inhibitors shows potential anticancer effects in LUAD cell lines, demonstrating the potential for the development potential pharmacological therapeutics for lung cancer by targeting PDE10A. Finally, we further identified that p.Pro360Ala on PDE10A may promote tumor cell growth. For instance, a colony formation assay showed that p.Pro360Ala on PDE10A significantly increased lung cancer cell growth compared with the wild-type and control groups, a finding suggestive of a potential oncogenic role. Since p.Pro360Ala is located at PDE10A allosteric binding site with druggability, it may represent an original targeted strategy in future pre/clinical studies by inhibiting the allosteric disregulation to PDE10A in lung cancer.

In this study, we revealed that the deleterious mutations identified in cancer genomes were more significantly enriched at known allosteric sites derived from protein X-ray structure data than tolerated mutations in proteins. Furthermore, the enrichment of deleterious variants could be of equal significance in potential allosteric sites predicted by the effect of ligand binding on protein dynamics, which will improve the identification of new allosteric sites. To validate the view, a widely used server in the allosteric field, SPACER,[67,68] was used to predict the most potential allosteric sites via binding leverage parameter. As a result, 40 allosteric sites from the server were carefully selected and then used to analyze the normalized deleterious/tolerated variant rate using AlloDriver. The analysis showed that deleterious variants of proteins were enriched at these potential allosteric sites in comparison with tolerated ones ($p = 0.0225$, Wilcoxon test), suggesting the same conclusion as we found in known allosteric sites.

Inspired by such discoveries, AlloDriver may not only shed light on the innovative molecular mechanisms of tumorigenesis by perturbing protein allosteric regulation but also enable the identification of novel allosteric sites based on somatic hotspot regions. We found that the deleterious mutations identified in cancer genomes were more significantly enriched at protein allosteric sites than tolerated mutations in the study, supporting a potential to identify allosteric sites from somatic hotspots. It should also be noted that deleterious mutations identified in cancer genomes can be significantly enriched at protein orthosteric sites when compared to tolerated mutations, and there is no statistical difference ($p = 0.24$) for deleterious variants at allosteric sites against orthosteric sites. Thus, our method is suitable to identify potential allosteric sites when protein orthosteric sites are well known. Otherwise, it is challenge to distinguish allosteric sites from orthosteric sites in the prediction based on directly

examining somatic hotspots. Machine learning-based model by constructing gold-standard negative and positive allosteric sites quantified by functional impact scores (e.g., SIFT and PolyPhen-2 scores) as descriptors may provide an alternative way to infer allosteric sites from somatic hotspots. This can be expanded in our future studies.

AlloDriver focused only on missense mutations that alter allosteric sites by single amino acid substitution by excluding other types of important mutations, such as nonsense mutations (stop codons), insertions/deletions (indels), or gene fusion. To reveal the effect of early stop codons, we systematically investigated nonsense mutations (stop codons) collected from approximately 7,000 matched tumor-normal samples at the experimentally validated allosteric sites for 74 allosteric proteins. In total, we found that among 74 allosteric proteins, 61 proteins had 474 nonsense variants and 40 of them located at allosteric sites. The mutational load for the nonsense variants (8.44% = 40/474) at the protein allosteric sites was significantly lower than that for the missense variants (21.0% = 935/4,451, $p = 1.8 \times 10^{-12}$, Fisher's exact test). The low distribution of stop codons at allosteric sites may result from the inherent feature of allosteric regulation. Allosteric regulation occurs through binding of a modulator at allosteric site to engender a conformational change that affects function at orthosteric site, and the coupling between allosteric site and orthosteric site are dependent heavily on protein dynamics supported by the scaffold of functional protein.[10] Nonsense variants of early stop codons result in various truncated proteins devoid of structure integrity, leading to the break of scaffold basis for most of allosteric function. For example, the truncated Abelson tyrosine kinase without SH2 and SH3 domains disabled the global allosteric regulation triggered by inhibitor GNF-5 at the allosteric site of kinase domain.[69] Therefore, the location of nonsense variants in allosteric proteins may evolutionarily occur everywhere instead of preferring to allosteric sites.

Supplemental Data

Supplemental Data include two figures and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.09.020.

Acknowledgments

This work was partially supported by the National High-tech R&D Program of China (863 Program) (2015AA020108) and the National Natural Science Foundation of China (81322046, 81473137, and 81302698) to J.Z., the National Natural Science Foundation of China (81573020) to F.C., and the NIH (R01LM011177) to Z.Z. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Received: May 27, 2016
Accepted: September 27, 2016
Published: December 8, 2016

Web Resources

ASD, http://mdl.shsmu.edu.cn/ASD
COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/
Ensemble BioMart database, http://useast.ensembl.org/Multi/Search/Results
GenBank, http://www.ncbi.nlm.nih.gov/genbank/
NCBI, http://www.ncbi.nlm.nih.gov/
NW-align, http://zhanglab.ccmb.med.umich.edu/NW-align/
OMIM, http://www.omim.org/
PDBSWS, http://bioinf.org.uk/pdbsws/
R statistical software (v.3.1.2), http://www.r-project.org/
RCSB Protein Data Bank, http://www.rcsb.org/pdb/home/home.do
Sanger website, ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl
The Cancer Genome Atlas, http://cancergenome.nih.gov/
UniProt ID mapping tool, http://www.uniprot.org/uploadlists/

References

1. Siegel, R.L., Miller, K.D., and Jemal, A. (2015). Cancer statistics, 2015. CA Cancer J. Clin. *65*, 5–29.
2. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.; and Cancer Genome Atlas Research Network (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.
3. Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S., et al.; International Cancer Genome Consortium (2010). International network of cancer genome projects. Nature *464*, 993–998.
4. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.
5. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.
6. Cheng, F., Zhao, J., and Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Brief. Bioinform. *17*, 642–656.
7. Vuong, H., Cheng, F., Lin, C.C., and Zhao, Z. (2014). Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach. Genome Med. *6*, 81–95.
8. De Smet, F., Christopoulos, A., and Carmeliet, P. (2014). Allosteric targeting of receptor tyrosine kinases. Nat. Biotechnol. *32*, 1113–1120.
9. Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000). Protein function in the post-genomic era. Nature *405*, 823–826.
10. Nussinov, R., and Tsai, C.J. (2013). Allostery in disease and in drug discovery. Cell *153*, 293–305.
11. Pe'er, D., and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. Cell *144*, 864–873.
12. Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. Semin. Cell Dev. Biol. *20*, 869–876.

13. Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., et al. (2009). Edgetic perturbation models of human inherited disorders. Mol. Syst. Biol. *5*, 321–330.
14. Cheng, F., Liu, C., Lin, C.C., Zhao, J., Jia, P., Li, W.H., and Zhao, Z. (2015). A Gene Gravity Model for the Evolution of Cancer Genomes: A Study of 3,000 Cancer Genomes across 9 Cancer Types. PLoS Comput. Biol. *11*, e1004497.
15. Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., Wang, Q., Shi, T., Zhao, Y., Wang, Y., et al. (2011). ASD: a comprehensive database of allosteric proteins and modulators. Nucleic Acids Res. *39*, D663–D669.
16. Huang, Z., Mou, L., Shen, Q., Lu, S., Li, C., Liu, X., Wang, G., Li, S., Geng, L., Liu, Y., et al. (2014). ASD v2.0: updated content and novel features focusing on allosteric regulation. Nucleic Acids Res. *42*, D510–D516.
17. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
18. UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.
19. Martin, A.C. (2005). Mapping PDB chains to UniProtKB entries. Bioinformatics *21*, 4297–4301.
20. Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics *10*, 168–179.
21. Schmidtke, P., Bidon-Chanal, A., Luque, F.J., and Barril, X. (2011). MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics *27*, 3276–3285.
22. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; and ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.
23. Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962.
24. Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.
25. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164–e171.
26. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.
27. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
28. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. *21*, 577–581.
29. Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., Biankin, A.V., Hautaniemi, S., and Wu, J. (2012). PINA v2.0: mining interactome modules. Nucleic Acids Res. *40*, D862–D865.
30. Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E., Brinkman, F.S., and Lynn, D.J. (2013). InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. Nucleic Acids Res. *41*, D1228–D1233.
31. Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. Nucleic Acids Res. *39*, D261–D267.
32. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database–2009 update. Nucleic Acids Res. *37*, D767–D772.
33. Newman, R.H., Hu, J., Rho, H.S., Xie, Z., Woodard, C., Neiswinger, J., Cooper, C., Shirley, M., Clark, H.M., Hu, S., et al. (2013). Construction of human activity-based phosphorylation networks. Mol. Syst. Biol. *9*, 655–666.
34. Hu, J., Rho, H.S., Newman, R.H., Zhang, J., Zhu, H., and Qian, J. (2014). PhosphoNetworks: a database for human phosphorylation networks. Bioinformatics *30*, 141–142.
35. Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. *40*, D261–D270.
36. Cheng, F., Jia, P., Wang, Q., Lin, C.C., Li, W.H., and Zhao, Z. (2014). Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. Mol. Biol. Evol. *31*, 2156–2169.
37. Cheng, F., Jia, P., Wang, Q., and Zhao, Z. (2014). Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. Oncotarget *5*, 3697–3710.
38. Benita, Y., Cao, Z., Giallourakis, C., Li, C., Gardet, A., and Xavier, R.J. (2010). Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. Blood *115*, 5376–5384.
39. Cheng, F., Zhao, J., Fooksa, M., and Zhao, Z. (2016). A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. J. Am. Med. Inform. Assoc. *23*, 681–691.
40. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517.
41. Zhang, M., Zhu, C., Jacomy, A., Lu, L.J., and Jegga, A.G. (2011). The orphan disease networks. Am. J. Hum. Genet. *88*, 755–766.
42. Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. *8*, e1002503.
43. Hirsh, A.E., Fraser, H.B., and Wall, D.P. (2005). Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol. Biol. Evol. *22*, 174–177.
44. Bezginov, A., Clark, G.W., Charlebois, R.L., Dar, V.U., and Tillier, E.R. (2013). Coevolution reveals a network of human

proteins originating with multicellularity. Mol. Biol. Evol. *30*, 332–346.
45. Capra, J.A., Williams, A.G., and Pollard, K.S. (2012). ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput. Biol. *8*, e1002567.
46. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.
47. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323–339.
48. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Methodol. *57*, 289–300.
49. Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. Stat. Med. *9*, 811–818.
50. Zhao, G.C., Sun, M., Wilde, S.A., and Li, S.Z. (2004). A Paleo-Mesoproterozoic supercontinent: assembly, growth and breakup. Earth Sci. Rev. *67*, 91–123.
51. Wenthur, C.J., Gentry, P.R., Mathews, T.P., and Lindsley, C.W. (2014). Drugs for allosteric sites on receptors. Annu. Rev. Pharmacol. Toxicol. *54*, 165–184.
52. Tzoneva, G., Perez-Garcia, A., Carpenter, Z., Khiabanian, H., Tosello, V., Allegretta, M., Paietta, E., Racevskis, J., Rowe, J.M., Tallman, M.S., et al. (2013). Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat. Med. *19*, 368–371.
53. Zhang, L., Hu, A., Yuan, H., Cui, L., Miao, G., Yang, X., Wang, L., Liu, J., Liu, X., Wang, S., et al. (2008). A missense mutation in the CHRM2 gene is associated with familial dilated cardiomyopathy. Circ. Res. *102*, 1426–1432.
54. Cheng, F., Li, W., Zhou, Y., Li, J., Shen, J., Lee, P.W., and Tang, Y. (2013). Prediction of human genes and diseases targeted by xenobiotics using predictive toxicogenomic-derived models (PTDMs). Mol. Biosyst. *9*, 1316–1325.
55. Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., Bottomley, W., et al. (2002). Mutations of the BRAF gene in human cancer. Nature *417*, 949–954.
56. Hobbs, G.A., Der, C.J., and Rossman, K.L. (2016). RAS isoforms and mutations in cancer at a glance. J. Cell Sci. *129*, 1287–1292.
57. Carpten, J.D., Faber, A.L., Horn, C., Donoho, G.P., Briggs, S.L., Robbins, C.M., Hostetter, G., Boguslawski, S., Moses, T.Y., Savage, S., et al. (2007). A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature *448*, 439–444.
58. Yi, K.H., Axtmayer, J., Gustin, J.P., Rajpurohit, A., and Lauring, J. (2013). Functional analysis of non-hotspot AKT1 mutants found in human breast cancers identifies novel driver mutations: implications for personalized medicine. Oncotarget *4*, 29–34.
59. Zhao, Q., Gu, X., Zhang, C., Lu, Q., Chen, H., and Xu, L. (2015). Blocking M2 muscarinic receptor signaling inhibits tumor growth and reverses epithelial-mesenchymal transition (EMT) in non-small cell lung cancer (NSCLC). Cancer Biol. Ther. *16*, 634–643.
60. Slattery, M.L., Lundgreen, A., and Wolff, R.K. (2012). MAP kinase genes and colon and rectal cancer. Carcinogenesis *33*, 2398–2408.
61. Chen, X., Wang, S., Wu, N., and Yang, C.S. (2004). Leukotriene A4 hydrolase as a target for cancer prevention and therapy. Curr. Cancer Drug Targets *4*, 267–283.
62. Visakorpi, T., Hyytinen, E., Koivisto, P., Tanner, M., Keinänen, R., Palmberg, C., Palotie, A., Tammela, T., Isola, J., and Kallioniemi, O.P. (1995). In vivo amplification of the androgen receptor gene and progression of human prostate cancer. Nat. Genet. *9*, 401–406.
63. Fujishige, K., Kotera, J., Michibata, H., Yuasa, K., Takebayashi, S., Okumura, K., and Omori, K. (1999). Cloning and characterization of a novel human phosphodiesterase that hydrolyzes both cAMP and cGMP (PDE10A). J. Biol. Chem. *274*, 18438–18445.
64. Handa, N., Mizohata, E., Kishishita, S., Toyama, M., Morita, S., Uchikubo-Kamo, T., Akasaka, R., Omori, K., Kotera, J., Terada, T., et al. (2008). Crystal structure of the GAF-B domain from human phosphodiesterase 10A complexed with its ligand, cAMP. J. Biol. Chem. *283*, 19657–19664.
65. Zhao, J., Cheng, F., Wang, Y., Arteaga, C.L., and Zhao, Z. (2016). Systematic prioritization of druggable mutations in ~5000 genomes across 16 cancer types using a structural genomics-based approach. Mol. Cell. Proteomics *15*, 642–656.
66. Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. USA *112*, E5486–E5495.
67. Mitternacht, S., and Berezovsky, I.N. (2011). Binding leverage as a molecular basis for allosteric regulation. PLoS Comput. Biol. *7*, e1002148.
68. Goncearenco, A., Mitternacht, S., Yong, T., Eisenhaber, B., Eisenhaber, F., and Berezovsky, I.N. (2013). SPACER: Server for predicting allosteric communication and effects of regulation. Nucleic Acids Res. *41*, W266–72.
69. Skora, L., Mestan, J., Fabbro, D., Jahnke, W., and Grzesiek, S. (2013). NMR reveals the allosteric opening and closing of Abelson tyrosine kinase by ATP-site and myristoyl pocket inhibitors. Proc. Natl. Acad. Sci. USA *110*, E4437–E4445.